On the Identification of a Class of Linear Models

Jin Tian
Department of Computer Science
Iowa State University
Ames, IA 50011
jtian@cs.iastate.edu

Abstract

This paper deals with the problem of identifying direct causal effects in recursive linear structural equation models. The paper provides a procedure for solving the identification problem in a special class of models.

Introduction

Structural equation models (SEMs) have dominated causal reasoning in the social sciences and economics, in which interactions among variables are usually assumed to be linear (Duncan 1975; Bollen 1989). This paper deals with one fundamental problem in SEMs, accessing the strength of linear cause-effect relationships from a combination of observational data and model structures.

The problem has been under study for half a century, primarily by econometricians and social scientists, under the name "The Identification Problem"(Fisher 1966). Although many algebraic or graphical methods have been developed, the problem is still far from being solved. In other words, we do not have a necessary and sufficient criterion for deciding whether a causal effect can be computed from observed data. Most available methods are sufficient criteria which are applicable only when certain restricted conditions are met.

In this paper, we show that the identification problem is solved in a special class of SEMs. We present a procedure that will decide whether each parameter in the model is identified or not and, if the answer is positive, the procedure will express the parameter in terms of observed covariances.

We begin with an introduction to SEMs and the identification problem, and give a brief review to previous work before presenting our results.

Linear SEMs and Identification

A linear SEM over a set of random variables $V = \{V_1, \ldots, V_n\}$ is given by a set of structural equations of the form

$$V_j = \sum_i c_{ji} V_i + \epsilon_j, \quad j = 1, \ldots, n, \tag{1}$$

where the summation is over the variables in V judged to be immediate causes of V_j. c_{ji}, called a *path coefficient*, quantifies the direct causal influence of V_i on V_j, and is also called a *direct effect*. ϵ_j's represent "error" terms and are assumed to have normal distribution. In this paper we consider recursive models and assume that the summation in Eq. (1) is for $i < j$, that is, $c_{ji} = 0$ for $i \geq j$. The set of variables (and the corresponding structural equations) are considered to be ordered as $V_1 < V_2 < \ldots < V_n$. We denote the covariances between observed variables $\sigma_{ij} = Cov(V_i, V_j)$, and between error terms $\psi_{ij} = Cov(\epsilon_i, \epsilon_j)$. We denote the following matrices, $\Sigma = [\sigma_{ij}]$, $\Psi = [\psi_{ij}]$, and $C = [c_{ij}]$. Without loss of generality, the model is assumed to be standardized such that each variable V_j has zero mean and variance 1.

$X = \epsilon_x$
$W = aX + \epsilon_w$
$Z = bX + \epsilon_z$
$Y = cW + dZ + \epsilon_y$
$Cov(\epsilon_x, \epsilon_z) \neq 0$
$Cov(\epsilon_w, \epsilon_y) \neq 0$

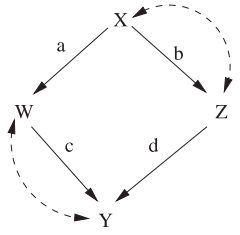


Figure 1: A linear SEM.

The structural assumptions encoded in the model are the zero path coefficient c_{ji}'s and zero error covariance ψ_{ij}'s. The model structure can be represented by a directed acyclic graph (DAG) G with (dashed) bidirected edges, called a *causal diagram* (or *path diagram*), as follows: the nodes of G are the variables $V_1, \ldots, V_n$; there is a directed edge from V_i to V_j in G if V_i appears in the structural equation for V_j, that is, $c_{ji} \neq 0$; there is a bidirected edge between V_i and V_j if the error terms ϵ_i and ϵ_j have non-zero correlation ($\psi_{ij} \neq 0$). Figure 1 shows a simple SEM and the corresponding causal diagram in which each directed edge is annotated by the corresponding path coefficient.

The parameters of the model are the non-zero entries in the matrices C and Ψ. Fixing the model structure and given parameters C and Ψ, the covariance matrix Σ is given by (see, for example, (Bollen 1989))

$$\Sigma = (I - C)^{-1} \Psi (I - C)^{T^{-1}}. \tag{2}$$

Conversely, in the identification problem, given the structure of a model, one attempts to solve for C in terms of the given

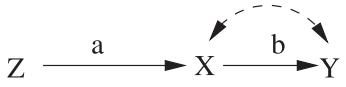


Figure 2: A typical instrumental variable

covariance Σ. If Eq. (2) gives a unique solution to some path coefficient c_{ji}, independent of the (unobserved) error correlations Ψ, that path coefficient is said to be *identified*. In other words, the *identification problem* is that whether a path coefficient is determined uniquely from the covariance matrix Σ given the causal diagram. If every parameter of the model is identified, then *the model is identified*. Note that the identifiability conditions we seek involve the structure of the model alone, not particular numerical values of parameters, that is, we insist on having *identification almost everywhere*, allowing for pathological exceptions (see, for example, (Brito & Pearl 2002a) for formal definition of identification almost everywhere).

Previous Work

Many methods have been developed for deciding whether a specific parameter or a model is identifiable. For example, the well-known instrumental variable (IV) method (Bowden & Turkington 1984) require search for variables (called *instruments*) that are uncorrelated with the error terms in specific equations. A typical configuration of the IV method is shown in Fig. 2, in which Z serves as an instrument for identifying the causal effect b as

$$b = \sigma_{ZY}/\sigma_{ZX}. \tag{3}$$

Traditional approaches are based on algebraic manipulation of the structural equations (Fisher 1966; Bekker, Merckens, & Wansbeek 1994; Rigdon 1995). Recently graphical approaches for identifying linear causal effects have been developed, and some sufficient graphical conditions were established (McDonald 1997; Pearl 1998; Spirtes *et al.* 1998; Pearl 2000; Tian 2004). The applications of these methods are limited in scope, and typically some special conditions have to be met for these methods to be applicable.

One principled approach for the identification problem is to write Eq.(2) for each term σ_{ij} of Σ using Wright's method of path coefficients (Wright 1934). Wright's equations consist of equating the (standardized) covariance σ_{ij} with the sum of products of parameters (c_{ji}'s and ψ_{ji}'s) along all *unblocked paths* between V_i and V_j. A path is *unblocked* if there is no node X such that both edges connected to X in the path have an arrow at X ($\rightarrow X \leftarrow$). A path coefficient c_{ij} is identified if and only if Wright's equations give a unique solution to c_{ij}, independent of error correlations. For example, the Wright's equations for the model in Fig. 2 are

$$\begin{aligned} \sigma_{ZX} &= a \\ \sigma_{ZY} &= ab \\ \sigma_{XY} &= b + \psi_{XY} \end{aligned} \tag{4}$$

Based on Wright's equations, a number of sufficient graphical criteria for *model identification* have been developed (Brito & Pearl 2002c; 2002b; 2006), which establish conditions for *all* the parameters in the model to be identified.

Recently, another principled approach for the identification problem is presented in (Tian 2005), which is similar to Wright's method but is based on exploiting partial regression coefficients. In this paper, we will use the partial regression coefficients method to solve the identifiability problem in a special class of SEMs, determining whether each individual parameter in the model is identifiable or not. First we introduce the method.

Partial regression equations

For a set $S \subseteq V$, let $\beta_{ij.S}$ denote the *partial regression coefficient* which represents the coefficient of V_j in the linear regression of V_i on V_j and S. (Note that the order of the subscripts in $\beta_{ij.S}$ is essential.) Partial regression coefficients can be expressed in terms of covariance matrices as follows (Cramer 1946):

$$\beta_{ij.S} = \frac{\Sigma_{V_iV_j} - \Sigma_{V_iS}^T \Sigma_{SS}^{-1} \Sigma_{V_jS}}{\Sigma_{V_jV_j} - \Sigma_{V_jS}^T \Sigma_{SS}^{-1} \Sigma_{V_jS}}, \tag{5}$$

where Σ_{SS} etc. represent covariance matrices over corresponding variables.

Let S_{jk} denote a set

$$S_{jk} = \{V_1, \ldots, V_{j-1}\} \setminus \{V_k\}. \tag{6}$$

(Tian 2005) derived an expression for the partial regression coefficient $\beta_{jk.S_{jk}}$, for each pair of variables $V_k < V_j$, in terms of the model parameters (path coefficients and error covariances) given by

$$\begin{aligned} \beta_{jk.S_{jk}} = c_{jk} + \alpha_{jk} - \sum_{k+1 \le l \le j-1} \beta_{lk.S_{lk}} \alpha_{jl}, \\ j = 2, \ldots, n,\ k = 1, \ldots, j-1, \end{aligned} \tag{7}$$

where α_{jk}'s are defined during the process of "orthogonalizing" the set of error terms to obtain a new set of error terms $\{\epsilon'_1, \ldots, \epsilon'_n\}$ that are mutually orthogonal in the sense that

$$Cov(\epsilon'_i, \epsilon'_j) = 0, \quad \text{for } i \neq j. \tag{8}$$

The Gram-Schmidt orthogonalization process proceeds recursively as follows. We set

$$\epsilon'_1 = \epsilon_1 \tag{9}$$

For $j = 2, \ldots, n$, we set

$$\epsilon'_j = \epsilon_j - \sum_{k=1}^{j-1} \alpha_{jk} \epsilon'_k \tag{10}$$

in which

$$\alpha_{jk} = \frac{Cov(\epsilon_j, \epsilon'_k)}{Cov(\epsilon'_k, \epsilon'_k)}. \tag{11}$$

The set of equations given by (7) are called the *partial regression equations*. As an example, the partial regression

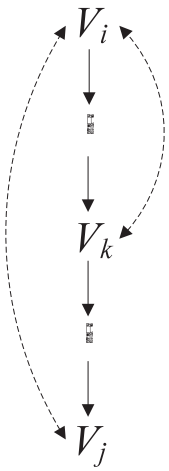


Figure 3: A P-structure

equations for the model shown in Figure 1 are given by

$$\beta_{WX} = a \tag{12}$$
$$\beta_{ZW.X} = 0 \tag{13}$$
$$\beta_{ZX.W} = b + \alpha_{ZX} \tag{14}$$
$$\beta_{YZ.WX} = d \tag{15}$$
$$\beta_{YW.XZ} = c + \alpha_{YW} \tag{16}$$
$$\beta_{YX.WZ} = -\beta_{WX}\alpha_{YW} \tag{17}$$

which happen to be linear with respect to all the parameters. It is not difficult to solve these equations to obtain that the path coefficients a, d, and c are identified.

Given the model structure (represented by zero path coefficients and zero error correlations), some of the c_{jk}'s and α_{jk}'s will be set to zero in Eq. (7), and we can solve the identifiability problem by solving Eq. (7) for c_{jk}'s in terms of the partial regression coefficients. This provides a principled method for solving the identifiability problem. A path coefficient c_{ij} is identified if and only if the set of partial regression equations give a unique solution to c_{ij}, independent of error correlations.

The partial regression equations are linear with respect to path coefficient c_{jk}'s and parameter α_{jk}'s, but may not be linear with respect to ψ_{ij}'s. α_{jk}'s are nonlinear functions of ψ_{ij}'s and may not be independent with each other. In this paper, we will identify a class of SEMs in which we can treat α_{jk}'s as independent free parameters and thus for this class of SEMs the partial regression equations become linear equations.

P-structure-Free SEMs

Definition 1 (P-structure) *For $j > k > i$, if there is a bidirected edge between V_j and V_i, and a bidirected edge between V_i and V_k, then we say that there is a* P-structure *in the causal diagram under the variable order $V_1 < \ldots < V_n$ (see Fig. 3). Equivalently, in terms of model parameters, we say that for $j > k > i$, if $\psi_{ji} \neq 0$ and $\psi_{ki} \neq 0$, then there is a P-structure in the SEM.*

This definition of P-structure depends on the order of the variables. In general we could rearrange the order of the variables as far as they are consistent with the topological order of the causal diagram. For example, the variables in Figure 1 can be ordered as $X < W < Z < Y$ or as $X < Z < W < Y$. It is possible that there is a P-structure in one order of the variables but not in another.

Definition 2 (P-structure-free Model) *We will say that a SEM (or causal diagram) is* P-structure free *if there exists an ordering of the variables that is consistent with the topological order of the causal diagram such that there is no P-structure in the causal diagram under this order.*

In this paper we will consider P-structure-free SEMs and assume that the variables are ordered $V_1 < \ldots < V_n$ such that there is no P-structure under this order.

First we show that in a P-structure-free SEM, α_{jk}'s can be treated as independent free parameters of the model.

Lemma 1 *In a P-structure-free SEM*

$$\alpha_{jk} = \frac{\psi_{jk}}{Cov(\epsilon'_k, \epsilon'_k)} \tag{18}$$

Proof sketch: Proof by induction. Assume Eq. (18) holds for all $\alpha_{j'k}$ such that $j' < j$, and for all $\alpha_{jk'}$ such that $k' < k$. Then

$$\begin{aligned}\alpha_{jk} &= \frac{Cov(\epsilon_j, \epsilon'_k)}{Cov(\epsilon'_k, \epsilon'_k)} \\ &= \frac{\psi_{jk} - c\sum_{l=1}^{k-1} \alpha_{kl} Cov(\epsilon_j, \epsilon'_l)}{Cov(\epsilon'_k, \epsilon'_k)} \\ &= \frac{\psi_{jk} - c\sum_{l=1}^{k-1} \psi_{kl}\psi_{jl}/Cov(\epsilon'_l, \epsilon'_l)}{Cov(\epsilon'_k, \epsilon'_k)} \\ &\quad \text{(induction hypothesis)}\end{aligned}$$

Now since there is no P-structure either ψ_{kl} or ψ_{jl} has to be zero. Therefore Eq. (18) holds. □

Corollary 1 *In a P-structure-free SEM $\alpha_{jk} = 0$ if and only if $\psi_{jk} = 0$. Graphically speaking, $\alpha_{jk} = 0$ if and only if there is no bidirected edge between V_j and V_k.*

From Lemma 1 it is clear that α_{jk}'s can be treated as independent parameters in place of ψ_{jk}'s in the set of equations given by (7). The identification problem is reduced to that solving Eq. (7) for c_{jk}'s in terms of the partial regression coefficients $\beta_{jk.S_{jk}}$'s, and a path coefficient c_{jk} is identified if and only if the set of partial regression equations (7) give a unique solution to c_{jk} that is independent of α_{jk}'s. The difficulty of solving these linear equations lies in that the coefficients of these equations, the partial regression coefficients, are not independent free parameters. The partial regression coefficients are related to each other in a complicated way, and it is difficult to decide the rank of the set of linear equations since it is not easy to determine whether certain expressions of partial regression coefficients will cancel out each other and become identically zero. To overcome this difficulty, next we show that the partial regression coefficients that appear in Eq. (7) can be expressed in terms of the free parameters c_{jk}'s and α_{jk}'s.

Figure 4: An active path between V_k and V_j given S_{jk}

We know that Wright's equations express covariance σ_{ij} in terms of the sum of products of parameters of the edges along paths between V_i and V_j. Here we will derive a similar result for the partial regression coefficients in a P-structure-free causal diagram. We assume that each directed edge $V_j \leftarrow V_k$ is associated with the path coefficient c_{jk} and each bidirected edge $V_j \leftrightarrow V_k$ is associated with the parameter α_{jk}. Next we show how a partial regression coefficient $\beta_{jk.S_{jk}}$ is related to the parameters along paths between V_j and V_k in a causal diagram. First, we define some graphical notations.

A *path* between two nodes X and Y in a causal diagram consists of a sequence of consecutive edges of any type (directed or bidirected). A non-endpoint node Z on a path is called a *collider* if two arrowheads on the path meet at Z, i.e. $\rightarrow Z \leftarrow$, $\leftrightarrow Z \leftrightarrow$, $\leftrightarrow Z \leftarrow$, $\rightarrow Z \leftrightarrow$; all other non-endpoint nodes on a path are *non-colliders*, i.e. $\leftarrow Z \rightarrow$, $\leftarrow Z \leftarrow$, $\rightarrow Z \rightarrow$, $\leftrightarrow Z \rightarrow$, $\leftarrow Z \leftrightarrow$.

Definition 3 (Active Path) *A path between two nodes X and Y is said to be* active *given a set of nodes Z if*
(i) every non-collider on the path is not in Z, and
(ii) every collider on the path is in Z.

Lemma 2 *In a P-structure-free SEM, every node V_l on an active path between V_k and V_j given S_{jk} where $k < j$ must be a collider and is ordered between V_k and V_j, $V_k < V_l < V_j$ (see Figure 4).*

The proof of Lemma 2 is ignored due to space constraints.

For a path p, let $T(p)$ represent the product of the parameters along path p. For example, let p be the path $V_1 \rightarrow V_2 \rightarrow V_6 \leftrightarrow V_8$ in Figure 5. Then $T(p) = c_{21}c_{62}\alpha_{86}$.

Lemma 3 *In a P-structure-free SEM,*

$$\beta_{jk.S_{jk}} = \sum_{p:\text{active path given } S_{jk}} (-1)^{|p|-1} T(p), \qquad (19)$$

in which the summation is over all the active paths between V_j and V_k given S_{jk} and $|p|$ represent the number of edges on p.

Proof idea: Proof by induction using Eq. (7) and Lemma 2. □

As a corollary of Lemma 3 we have that $\beta_{jk.S_{jk}} = 0$ if there is no active path between V_j and V_k given S_{jk}, which essentially says that $\beta_{jk.S_{jk}} = 0$ if V_j is conditionally independent of V_k given S_{jk} (see (Spirtes *et al.* 1998)).

Next, we show how to solve the set of partial regression equations given by Eq. (7) in a P-structure-free SEM.

Identifying P-structure-free SEMs

In a P-structure free SEM, to decide the identifiability of the path coefficients associated with a variable V_j, c_{jk}'s, all we need to do is to solve the $j-1$ equations in (7), $k = 1, \ldots, j-1$, for c_{jk}'s in terms of $\beta_{jk.S_{jk}}$'s, and c_{jk} is identified if and only if the set of equations give a unique solution to c_{jk}. Each of the equation expresses the active paths between V_j and a variable and we will name each equation after the corresponding variable as

$$(V_k) : \beta_{jk.S_{jk}} = c_{jk} + \alpha_{jk} - \sum_{k+1 \le l \le j-1} \beta_{lk.S_{lk}} \alpha_{jl}. \quad (20)$$

Consider the case that there is a directed edge from V_k to V_j, $V_k \rightarrow V_j$, in the causal diagram. The path coefficient c_{jk} only appears once in this $j-1$ equations, that is, in the equation (V_k). We can express c_{jk} in terms of α_{ji}'s,

$$c_{jk} = \beta_{jk.S_{jk}} - \alpha_{jk} + \sum_{k+1 \le l \le j-1} \beta_{lk.S_{lk}} \alpha_{jl}. \qquad (21)$$

Therefore c_{jk} is identifiable if none of the α_{ji}'s appears in this equation or all the α_{ji}'s appearing in the equation are identifiable.

Next, we consider the rest of equation (V_k)'s given by (20) where there is no directed edge from V_k to V_j and therefore $c_{jk} = 0$. Let Γ_j denote this set of linear equations, in which α_{ji}'s are the variables. In general Γ_j may have more equations than variables, or more variables than equations. A common approach to represent the structures of systems of equations is using bipartite graphs. A *bipartite graph* is an undirected graph $G = (N, E)$ in which the set of nodes N can be partitioned into two sets N_1 and N_2 such that all edges in E go between the two sets N_1 and N_2. A *matching* in a bipartite graph is a set of edges that do not share nodes. A matching is *maximal* if it is not properly contained in any other matching. A node X is *matched* by matching M if some edge in M is incident on X, otherwise X is *unmatched*. We use a bipartite graph to represent the relations between equations and variables as follows. Let each node in N_1 represent an equation, each node in N_2 represent a variable, and each edge in E represent that the variable appears in the corresponding equation. For example, Figure 6 shows the bipartite graph representation for the set of equations (25) to (28).

Let BG be the bipartite graph representing the set of equations Γ_j. Let M be a maximal matching in BG. Let Δ_j be the set of equations that are matched by the matching M and Δ'_j be the set of equations that are unmatched by M. The equations in Δ'_j are redundant considering only the structural information of the equation system. The number of equations in Δ_j is no more than the number of variables in Δ_j. For any set of equations S, let $|S|$ denote the number of equations in S, and let n_S denote the number of variables in S.

Definition 4 (Non-over-constrained System) *A system of equations S is non-over-constrained if any subset of $k \le |S|$ equations of S contains at least k different variables.*

Lemma 4 *Δ_j is non-over-constrained.*

Proof: Any subset of k equations of Δ contains at least the k variables that match the subset of equations in the matching M. □

One approach for solving a non-over-constrained system is using Simon's causal ordering algorithm (Simon 1953), extended in (Lu, Druzdzel, & Leong 2000). The algorithm works by successively solving self-contained set of equations. A non-over-constrained set of equations S is *self-contained* if $|S| = n_S$. A self-contained set is *minimal* if it does not contain any self-contained subsets.

Next we present the causal ordering algorithm as described in (Lu, Druzdzel, & Leong 2000). The algorithm starts with *identifying* the mininal self-contained subsets in the input system K. Then we *solve* these subsets, remove the equations in those subsets from K, and *substitute* the values of solved variables into remaining equations. The remaining set of equations is called the *derived system*. We keep applying identifying, solving, and substitution operations on derived system until either the derived system D is empty, which means that K is self-contained, or there are no more self-contained subsets that can be identified, which means that the set of remaining variables in D can not be solved and D is called a *derived strictly under-constrained subsets*.

The following lemma shows that we can always solve a self-contained subset of Δ_j, and therefore the causal ordering algorithm can be used to solve Δ_j.

Lemma 5 *The set of equations in any self-contained subset A of Δ_j, or in a derived self-contained subset of Δ, are linear independent. That is, A can be solved for unique values of the set of variables in A.*

Proof idea: Let the set of equations be $\{(Z_1), \ldots, (Z_m)\}$. Assume that each equation (Z_i) matches the variable $\alpha_{V_j A_i}$. Then for every node Z_i, there is an active path p_i between Z_i and V_j that includes the edge $A_i \leftrightarrow V_j$ by Lemma 3. The determinant of the coefficient matrix of the set of equations will contain the term $\prod_i T(p_i)$. This term can not be cancelled out by any other term because any active path between Z_l and V_j that includes the edge $A_i \leftrightarrow V_j$ must contain an edge that does not appear in any of $p_1, \ldots, p_n$. Note that the proof is similar to the proof for Theorem 1 in (Brito & Pearl 2006). □

After applying the causal ordering algorithm to solve Δ_j, if the derived system D is empty, then every α_{jk} is identified. If the derived system D is a derived strictly under-constrained subsets, then the set of variables in D can not be identified. Therefore this procedure will tell which α_{ji}'s are identifiable, and which α_{ji}'s are not identifiable. Then the identifiability of the path coefficients c_{jk}'s can be determined by Eq. (21). Although the equations in Δ'_j are redundant in the sense that they are not useful for determining the identifiability of parameters, they lead to constraints on the covariance matrix implied by the SEM. Finally, to decide the identifiability of every path coefficients in the model, we apply this procedure for each j from 2 to n.

The following theorem state the conditions for the model identification.

Theorem 1 *A P-structure-free SEM is identified if and only if Δ_j is self-contained for every j.*

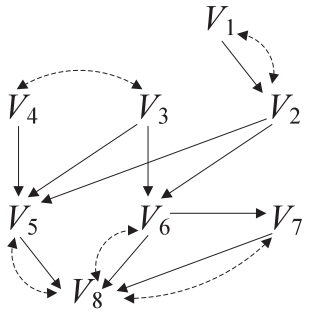


Figure 5: A SEM

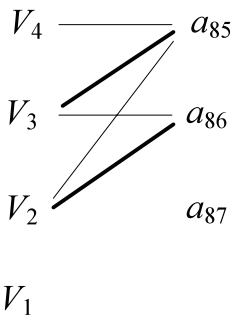


Figure 6: The bipartite graph representation of equations

An example

We illustrate the procedure we presented by applying it to the model given in Figure 5. Assume that we would like to decide the identifiability of the path coefficients associated with V_8. First we express c_{87}, c_{86}, c_{85} in terms of α_{87},α_{86},α_{85}:

$$c_{87} = \beta_{87.S_{87}} - \alpha_{87} \tag{22}$$
$$c_{86} = \beta_{86.S_{86}} - \alpha_{86} + \beta_{76.S_{76}}\alpha_{87} \tag{23}$$
$$c_{85} = \beta_{85.S_{85}} - \alpha_{85} \tag{24}$$

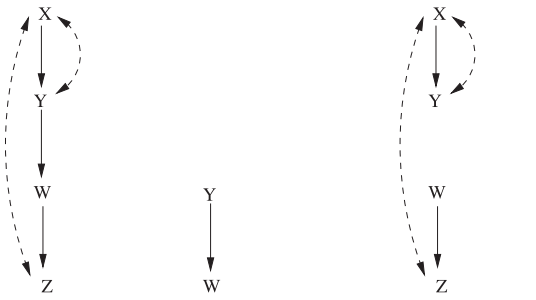
Then we give the set of equations corresponding to variables V_1, V_2, V_3, V_4:

$$V_4: \quad \beta_{84.S_{84}} = -\beta_{54.S_{54}}\alpha_{85} \tag{25}$$
$$V_3: \quad \beta_{83.S_{83}} = -\beta_{53.S_{53}}\alpha_{85} - \beta_{63.S_{63}}\alpha_{86} \tag{26}$$
$$V_2: \quad \beta_{82.S_{82}} = -\beta_{52.S_{52}}\alpha_{85} - \beta_{62.S_{62}}\alpha_{86} \tag{27}$$
$$V_1: \quad \beta_{81.S_{81}} = 0 \tag{28}$$

The bipartite graph representation of these set of equations is shown in Figure 6, which also shows a maximum matching in which (V_3) and (V_2) are matched. Equations (V_3) and (V_2) form a minimal self-contained set, and they can be solved to obtain a solution for α_{85} and α_{86}

$$\alpha_{85} = \frac{\beta_{82.S_{82}}\beta_{63.S_{63}} - \beta_{83.S_{83}}\beta_{62.S_{62}}}{\beta_{53.S_{53}}\beta_{62.S_{62}} - \beta_{52.S_{52}}\beta_{63.S_{63}}} \tag{29}$$
$$\alpha_{86} = \frac{\beta_{83.S_{83}}\beta_{52.S_{52}} - \beta_{82.S_{82}}\beta_{53.S_{53}}}{\beta_{53.S_{53}}\beta_{62.S_{62}} - \beta_{52.S_{52}}\beta_{63.S_{63}}} \tag{30}$$

α_{87} does not appear in any equations and therefore is not identifiable. Equations (V_1) and (V_4) are unmatched and



(a) G (b) $G(\{W\})$ (c) $G(\{X, Y, Z\})$

Figure 7: A SEM model

lead to constraints implied by the model. More specifically, equation (V_1) represents the conditional independence constraints that V_8 is conditionally independent of V_1 given all other variables. Substituting the value of α_{85} into equation (V_4) leads to the following constraints

$$\beta_{84.S_{84}} = -\beta_{54.S_{54}} \frac{\beta_{82.S_{82}}\beta_{63.S_{63}} - \beta_{83.S_{83}}\beta_{62.S_{62}}}{\beta_{53.S_{53}}\beta_{62.S_{62}} - \beta_{52.S_{52}}\beta_{63.S_{63}}} \quad (31)$$

Finally, substituting the solutions for α_{85} and α_{86} into Eqs. (22) to (24) we conclude that c_{85} is identified as

$$c_{85} = \beta_{85.S_{85}} - \frac{\beta_{82.S_{82}}\beta_{63.S_{63}} - \beta_{83.S_{83}}\beta_{62.S_{62}}}{\beta_{53.S_{53}}\beta_{62.S_{62}} - \beta_{52.S_{52}}\beta_{63.S_{63}}} \quad (32)$$

and that c_{86} and c_{87} are not identifiable.

Conclusion and Discussions

The identification problem has been a long standing problem in the applications of linear SEMs. Given a SEM, we would like to know which parameters in the model are uniquely determined by the observed covariances and which parameters are not, and we would like to know what constraints are implied by the model structure on the covariance matrix. In this paper, we provide a procedure for answering these questions in a special class of SEMs.

The applications of this result may be broadened by combining it with a model decomposition technique given in (Tian 2005), which shows that a model can be decomposed into a set of submodels such that the identification problem can be solved independently in each submodel. It is possible that a model which is not P-structure-free may be decomposed into submodels all of which are P-structure-free. For example, the model in Figure 7(a) is not P-structure-free. However (Tian 2005) shows that the identification problem can be solved independently in the two models in Figure 7(b) and (c). Now the model in Figure 7(c) is P-structure-free under the order $W < Z < X < Y$. Hence, we conclude that the identification problem in the model in Figure 7(a) can be solved by the procedure given in this paper.

Acknowledgments

This research was partly supported by NSF grant IIS-0347846.

References

Bekker, P.; Merckens, A.; and Wansbeek, T. 1994. *Identification, equivalent models, and computer algebra*. Academic.

Bollen, K. 1989. *Structural Equations with Latent Variables*. New York: John Wiley.

Bowden, R., and Turkington, D. 1984. *Instrumental Variables*. Cambridge, England: Cambridge University Press.

Brito, C., and Pearl, J. 2002a. Generalized instrumental variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 85–93. San Francisco, CA: Morgan Kaufmann.

Brito, C., and Pearl, J. 2002b. A graphical criterion for the identification of causal effects in linear models. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*, 533–538. Menlo Park, CA: AAAI Press/The MIT Press.

Brito, C., and Pearl, J. 2002c. A new identification condition for recursive models with correlated errors. *Structural Equation Modelling* 9(4):459–474.

Brito, C., and Pearl, J. 2006. Graphical condition for identification in recursive sem. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press.

Cramer, H. 1946. *Mathematical Methods of Statistics*. Princeton, N.J: Princeton University Press.

Duncan, O. 1975. *Introduction to Structural Equation Models*. New York: Academic Press.

Fisher, F. 1966. *The Identification Problem in Econometrics*. McGraw-Hill.

Lu, T.-C.; Druzdzel, M. J.; and Leong, T. Y. 2000. Causal mechanism-based model construction. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–2000)*, 353–362. San Francisco, CA: Morgan Kaufmann Publishers.

McDonald, R. 1997. Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research* 32(1):1–38.

Pearl, J. 1998. Graphs, causality, and structural equation models. *Socioligical Methods and Research* 27:226–284.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. NY: Cambridge University Press.

Rigdon, E. 1995. A necessary and suficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research* 30:359–383.

Simon, H. 1953. Causal ordering and identifiability. In Hood, W. C., and Koopmans, T., eds., *Studies in Econometric Method*. Wiley and Sons, Inc. 49–74.

Spirtes, P.; Richardson, T.; Meek, C.; Scheines, R.; and Glymour, C. 1998. Using path diagrams as a structural equation modeling tool. *Socioligical Methods and Research* 27:182–225.

Tian, J. 2004. Identifying linear causal effects. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 104–110. AAAI Press/The MIT Press.

Tian, J. 2005. Identifying direct causal effects in linear models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. AAAI Press/The MIT Press.

Wright, S. 1934. The method of path coefficients. *Ann. Math. Statist.* 5:161–215.